DOCUMENT RESUME

ED 310 119 TM 013 692

AUTHOR Kelderman, Henk
TITLE Item Bias Detection Using the Loglinear Rasch Model: Observed and Unobserved Subgroups. Research Report 86-2.
INSTITUTION Twente Univ., Enschede (Netherlands). Dept. of Education.
PUB DATE 86
NOTE 49p.; Also cited as Project Psychometric Aspects of Item Banking No. 3.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Foreign Countries; Higher Education; *Latent Trait Theory; Mathematical Models; Multiplication; Statistical Analysis; *Statistical Bias; *Test Bias; Testing Problems; Test Items; Undergraduate Students
IDENTIFIERS Contingency Tables; *Item Bias Detection; Item Parameters; Latent Class Models; Log Linear Models; Netherlands; *Rasch Model; Subgroups

ABSTRACT

A method is proposed for the detection of item bias with respect to observed or unobserved subgroups. The method uses quasi-loglinear models for the incomplete subgroup x test score x item 1 x ... x item k contingency table. If the subgroup membership is unknown, the models are the incomplete-latent-class models of S. J. Haberman (1979). The (conditional) Rasch model is formulated as a quasi-loglinear model. The parameters in this model that correspond to the main effects of the item responses are the conditional estimates of the parameters in the Rasch model. Item bias can then be tested by comparing the quasi-loglinear-Rasch model with models that contain parameters for the interaction of item responses and the subgroups. An example uses data from a test taken by 286 Dutch undergraduates who took a multiplication test using Roman numerals and numbers written out in Dutch. Some of the examinees had received training in multiplying Roman numerals. It was expected that Roman items would be biased, and the procedure confirmed this bias. Five tables present the models and study data. A 55-item list of references is included. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED310119

Item Bias Detection using the Loglinear Rasch Model

Observed and Unobserved Subgroups

Research Report 86-2

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. NELISSEN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

H. Kelderman

BEST COPY AVAILABLE

Division of Educational Measurement and Data Analysis

University of Twente

TM013692

Department of Education

Project Psychometric Aspects of Item Banking No. 3

Colophon

Typing : A. Burchartz, L. Padberg
Cover design : M. Driessen, AV-section, University of Twente
Printed by : Central Reproduction Department, University of Twente

Item Bias Detection using the Loglinear Rasch Model: Observed and Unobserved Subgroups

Henk Kelderman

Abstract

A method is proposed for the detection of item bias with respect to observed or unobserved subgroups. The method uses quasi-loglinear models for the incomplete subgroup x testscore x item 1 x ... x itek k contingency table. If subgroup membership is unknown the models are Haberman's incomplete-latent-class models.

The (conditional) Rasch model is formulated as a quasi-loglinear model. The parameters in this loglinear model, that correspond to the main effects of the item responses, are the conditional estimates of the parameters in the Rasch model. Item bias can then be tested by comparing the quasi-loglinear-Rasch model with models that contain parameters for the interaction of item responses and the subgroups.

Introduction

Educational or psychological tests are biased if the testscores of equally able test takers are systematically different between racial, ethnic, cultural etc. subgroups. Biased test scores may lead to unfair decisions or erroneous conclusions about individuals from particular subgroups. A test score is biased only if one or more of the test items are biased. A test item is biased if individuals with the same ability level from different subgroups have a different probability of a right response, i.e. the item has different difficulties in different subgroups. A test can be made fairer by deleting or improving the biased items.

Binet and Simon (1916; see Jensen 1980, p367) were already concerned with bias when they applied their test of general intelligence that was standardized on working class childeren to children of higher social status.

To assess bias some unbiased criterion measure of ability is needed. In some studies an external criterion for ability is at hand (e.g. Petersen and Novick, 1976). In most practical situations, however, no such external criterion is available and some criterion for ability internal to the test itself is used. Therefore most item bias detection techiques that are discussed in the literature use an internal criterion in some manner. The way in which this is done best distinguishes the methods from each other. Reviews are given by Osterlind (1983); Rudner Getson and Knight (1980); and Shepard, Camilli and Averill (1981). Handbooks on item

bias detection methods and research are Berk (1982) and Jensen (1980).

In the earlier item bias detection methods there is no explicit control of ability. For instance, in the analysis of covariance approach (Cardall & Coffman, 1965) transformed p-values are analyzed in a subgroup x items design. If there is a significant item by subgroup interaction, the item is considered biased. The analysis of variance assumption of equal cell variances is met by transforming the p-values by an arcsin transformation. Cleary and Hilton (1968), Hoepfner and Strickland (1972) and Jensen (1973) give further examples of this method.

The oldest and most popular item bias detection method is the transformed item difficulty method (Thurstone, 1925; Angoff, 1982; Angoff & Ford, 1973). It is conceptually very similar to the analysis of variance method, because it also studies the item x subgroup interaction of item difficulty. Angoff converted each p-value to a normal deviate (called delta's) by an inverse normal transformation. For all items delta values are compared between two subgroups by plotting these pairs of delta's in a bivariate graph. Angoff claims that the delta pairs for each item scatter around a straight line if the items are unbiased. If an item falls at some distance from the line this indicates an item x subgroup interaction. The item is then considered biased. Examples of practical application of this method are in Dorans (1982) and Donlon, Hicks and Wallmark (1980).

Both the analysis of variance method and the transformed item

difficulty method analyse subgroup x items interactions in item statistics on the subgroup level. Consequently, in these methods control for ability must be performed through correcting for differences between subgroup-level item statistics. This is logically unsatisfactory because, according to the definition of item bias, control for ability must be performed on the individual level. It is also unsatisfactory in practice. Hunter (1975) and Shepard Camilli and Williams (1985) show that when the items vary in difficulty and the distribution of ability is different in different subgroups, items x subgroups interactions can arise in perfectly unbiased tests.

A better way to control for ability is to use the raw score of the remaining test items as an estimate of ability. Item bias detection methods based on this idea, called chi-square methods, are proposed by Scheuneman (1979) and Mellenbergh (1982). Scheuneman uses data from an item response x subgroup x scoregroup contingency table to test the hypothesis that within each scoregroup the probabilities of a positive item response are the same for all subgroups. If the hypothesis is rejected the item are considered biased. Baker (1981) criticized Scheuneman's methods on the grounds that the distribution of the test statistic is unknown because Scheuneman used only the data from the positive responses. Camilli (1979) and Nungester (1977) (see Ironson 1982) proposed a test statistic based on both the correct and the incorrect item responses which is asympotically distributed as a chi-square. Mellenbergh (1982) modified Scheuneman's method so that it fits in

the general theory of loglineair and logit models for contingency tables. This yields a parametric model describing different types of bias which can also be tested by chi-square statistics.

Chi-square methods can detect item bias very well. Rudner, Getson and Knight (1980) show that Scheuneman's method can detect item bias in simulated data where the responses are generated form a three parameter logistic model with different slope and locations in different groups. Van der Flier, Mellenbergh, Adër and Wijn (1984) show that Mellenbergh's (1982) method works well in both empirical data and in simulated data generated by a certain three-parameter-normal ogive type model. Kok, Mellenbergh and van der Flier (1985) showed that the method also effectively detected experimentally induced item bias. Although in chi-square methods there is a better control for ability level than in the analysis of variance method and the transformed item difficulty method, taking ability as the number right score on the remaining items is rather informal and possibly inappropriate.

In item-response theory, ability is described by formal parameters. In these models the probability of an individual response to a certain item is explained by parameters describing the individual's ability and the item's difficulty. An item is considered biased if the item parameters are different for individuals with the same ability parameters from different subgroups.

Lord (1980) used Birnbaum's three parameter logistic model to detect biased items. The model contains three parameters: a lower

asymptote, a slope, and a location parameter of the item characteristic curve. The parameters are associated with guessing, discrimination, and difficulty, respectively. If one or more of tne estimated item parameters differ significantly from subgroup to subgroup, the item is considered biased.

Muthen and Lehman (1985) uses multiple group factor analysis of dichotomous variables to test the invariance of the parameters of the two-parameter-normal-ogive model over subgroups.

Durovic (1975) as well as Wright, Mead and Draba (1975) use the Rasch model to detect item bias. For each item the mean squared differences between the observed responses and expected probabilities of a correct response were computed and compared between two subgroups.

In this paper the loglinear formulation of the Rasch model (Kelderman, 1984) is used to test the invariance of item parameters over subgroups. If the difficulty parameters vary from subgroup to subgroup the item is considered biased. Subgroup membership may be observed or unobserved. In some practical situations, items may be expected to be biased for certain subgroups of individuals, but it is not known a-priori to which subgroup each of the individuals belongs. For example, for an item in an examination the probability of a correct response may be larger for a group of individuals with specific educational experiences than for individuals without that experience, or for an item in a mastery test the probability of a correct response may be larger for a subgroup of individuals having a different study strategy or for a subgroup of individuals having

a different cognitive strategy to solve the item, etc. In these examples, information on the individuals' subgroup membership may be difficult to observe or, as in the last example, the test behavior itself may be the natural indicator of subgroup membership. In this paper a loglinear Rasch model is formulated where item difficulty may also vary over subgroups that are not observed.

In what follows, the choice of the Rasch model to detect item bias is discussed. Quasi-loglinear models are formulated for test data and the Rasch model is formulated as one of them. Some alternative models are described to test various aspects of item bias with respect to known subgroups. The use of these tests is illustrated on a set of test data from Kok (1982) where item bias was introduced experimentally. Finally, corresponding latent class models for item bias with respect to unknown subgroups are described, and the effects of this bias is discussed.

Choice of model

The Rasch model describes the probability $P(X_j=x_j|\alpha)$ that an individual with parameter α gives a response X_j to item j $(j=1,\ldots,k)$, where X_j can take values $x_j = 0,1$ for a wrong (0) or a right (1) response:

$$(1) \qquad P(X_j=x_j|\alpha) = \exp(x_j(\alpha-\delta_j)) \;/\; (1+\exp(\alpha-\delta_j)) \;,$$

where δ_j (j=1,...,k) is a single item parameter describing the difficulty of item j. If this item parameter varies from subgroup to subgroup, the item is considered biased. Although the Rasch model is a rather simple model, its parsimony yields several virtues in using it to detect item bias.

Firstly, unlike the Birnbaum model if in the Rasch model item A has a larger item parameter value than item B, the probability of getting a correct solution on item A is always smaller than the probability of getting a correct solution on items B regardless of the examinee's ability level. Consequently, if the data fit the Rasch model, it makes sense to assert that item A is more difficult than item B. The item parameter value may therefore justifiably be interpreted as the item's difficulty (Rasch, 1966a), so that differences in item parameters between different subgroups can be interpreted as differences in item difficulty between subgroups. The dependence on the subgroups of the item parameters can then be analyzed to make a diagnosis of the item's flaws necessary to improve the item.

Secondly, in item bias detection studies we are interested in invariance of item parameters over subgroups and not in the individual person parameter values within each subgroup. It is therefore a desirable property of the Rasch model that the item parameters are inferentially separable from the person parameters. The Rasch model is an exponential family model wherein the simple number right score $T = X_1 + \ldots + X_k$ is a sufficient statistic for the person parameter α. Assuming local independence of the item

responses for a given value of α and after conditioning on the number right score taking the value t, the joint probability $P(X_1=x_1,\ldots,X_k=x_k|T=t)$ of the item responses $X_1, \ldots, X_k$ for a given score T=t becomes Rasch (1966b):

$$(2) \qquad P(X_1=x_1,\ldots,X_k=x_k|T=t)$$

$$= \exp(-x_1\delta_1\ldots-x_k\delta_k)/(\sum_{t=x_1+\ldots+x_k} \ldots \sum_{k} \exp(-x_1\delta_1\ldots-x_k\delta_k)) ,$$

By conditioning on the score, the nuisance parameter α has vanished (Rascn 1966b). In this paper the invariance over subgroups i (i=1,...,m) of the joint item response distributions for given values of T

$$(3) \qquad P_i(X_1=x_1,\ldots,X_k=x_k|T=t) = P(X_1=x_1,\ldots,X_k=x_k|T=t)$$

is tested to study item bias. According to model (2) any deviation of this invariance must be explained by differences in item difficulty between the subgroups. Note from (2) that the use of the Rasch model to study item bias is both an observed score method and a latent-trait-model method.

Thirdly, the conditional Rasch model can be formulated (Kelderman, 1984) as a quasi-loglinear contingency table model (Fienberg, 1972; Bishop, Fienberg & Holland, 1975). Model (2) is then equivalent to the hypothesis that the item responses and the score are quasi independent (Goodman, 1968) in the incomplete score

x item 1 x ... x item k contingency table. Incomplete table methodology can be used to formulate several hypotheses about item bias by specifying alternative quasi-loglinear models that contain various subgroup dependent parameters. Testing the conditional Rasch model against such models yields a test of the hypotneses.

Quasi-Loglinear-Models for the Incomplete Subgroup x Score x Item 1 x ... x Item k Table.

Let $f_{itx_1 \ldots x_1}$ be the number of individuals from subgroup i (i=1,...,m) with number right score T=t (t=0,1,...,k) and item scores $X_1=x_1, \ldots, X_k=x_k$ where $x_j = 1$ if item j (j=1,...,k) is answered correctly and $x_j = 0$ if item j is answered incorrectly. Since it is logically impossible to have a test score that is unequal to the number of correct item responses (excluding counting errors) the counts $f_{itx_1 \ldots x_k}$ are zero for $t \neq \sum_i x_i$. Table 1 shows the

table 1

subgroup x score x item 1 x ... x item 3 contingency table for subgroup i. Dashes denote cells that are logically or structurally zero cell. Contingency tables with structurally zero cells are called incomplete contingency tables.

Fienberg (1972; see also Bishop, Fienberg & Holland, 1975) presents a general theory for the statistical analysis of

incomplete multiway contingency tables by quasi-loglinear models. We apply Fienbergs theory to the analysis of the subgroup x score x item 1 x ... x item k contingency table to detect item bias.

Let $m_{itx_1 \ldots x_k}$ be the expected counts for the table under some model. If $t \neq x_1 + \ldots + x_k$ the expected counts are again structurally zero. If $t = x_1 + \ldots + x_k$, the expected counts are structurally nonzero and these counts are explained by a quasi-loglinear model. The saturated or fully specified model for the table is:

(4) $\ln m_{itx_1 \ldots x_k} =$

$$u + u_1(i) + u_2(t) + u_3(x_1) + \ldots + u_{(k+2)}(x_k)$$

$$+ u_{12}(it) + u_{13}(ix_1) + \ldots + u_{(k+1)(k+2)}(x_{k-1}x_k)$$

$$+ u_{123}(itx_1) + \ldots + u_{123\ldots(k+2)}(itx_1 \ldots x_k)$$

for $i = 1, \ldots, m$; $x_1 = 0,1$; $\ldots$; $x_k = 0,1$; $t = x_1 + \ldots + x_k$, where ln is the natural logarithm. Model (4) has constraints:

(5) $u_1(+) = u_2(+) = \ldots = u_{(k+2)}(+) = u_{12}(+t) = u_{12}(i+) =$
$= u_{13}(+x_1) = u_{13}(i+) = = u_{(k+1)(k+2)}(+x_k) =$
$= u_{(k+1)(k+2)}(x_{k-1}+) = \ldots = u_{123}(+tx_1) = u_{123}(i+x_1) =$
$= u_{123}(it+) = \ldots = u_{123\ldots(k+2)}(+tx_1 \ldots x_k) =$
$= u_{123\ldots(k+2)}(i+x_1 \ldots x_k) =$
$= u_{123\ldots(k+2)}(itx_1 \ldots x_{k-1}+) = 0.$

The u-terms in model (4) describe main effects and interaction effects of subgroup i, score t and item responses x_1, ..., x_k. The u-terms in expression (5) denote sums of parameters that occur in model (4) where a plus sign replacing an index indicates that the summation is over the replaced index. The constraints (5), however, are not sufficient to ensure that all parameters in model (4) are estimable. Additional constraints must be imposed to obtain a unique solution of the model parameters. These constraints will be discussed later.

Restrictive quasi-loglinear models are defined by setting u-terms in (4) equal to zero. The only models considered here will be hierarchical, i.e. whenever a particular u-term is set to zero, all its higher order relatives must also be set to zero.

The Rasch Model as a Quasi-Loglinear Model.

A restrictive quasi-loglinear model is

$$(6) \qquad \ln m_{itx_1 \ldots x_k} = u + u_1(i) + u_2(t) + u_{12}(it) + u_3(x_1) + \ldots + u_{k+2}(x_k)$$

with the constraints

$$(7) \qquad u_1(+) = u_2(+) = u_{12}(+t) = u_{12}(i+) = u_3(+) = \ldots = u_{k+2}(+) = 0$$

Model (6) can be obtained from the saturated quasi-loglinear model (4) by setting all interactions with and between item responses equal to zero.

If the subgroup and score are taken as fixed variables and the item responses are considered as random variables, model (6) is equivalent to the conditional Rasch model. In that case $m_{itx_1 \ldots x_k}$ is the conditional expected frequency of the response $X_1=x_1, \ldots, X_k=x_k$ for given subgroup i and score t. The conditional probability of response $X_1=x_1, \ldots, X_k=x_k$ for i and t can then be obtained from (6) by

$$(8) \qquad P_i(X_1=x_1,\ldots,X_k=x_k \mid T=t) = m_{itx_1\ldots x_k} / \sum_{x_1} \cdots \sum_{x_k} m_{itx_1\ldots x_k}$$
$$x_1+\ldots+x_k=t$$

$$= \exp(u_3(x_1)+\ldots+u_{k+2}(x_k)) / \sum_{x_1} \cdots \sum_{x_k} \exp(u_3(x_1)+\ldots+u_{k+2}(x_k)).$$
$$x_1+\ldots+x_k=t$$

Except for a reparametrization, model (8) is equivalent to model (2). In model (2) the effect $-x_j\delta_j$ of a response $X_j=x_j$ on item j is $-\delta_j$ for a correct response ($X_j=1$) and zero for an incorrect response ($X_j=0$), whereas in model (8) the effect of a correct response is $u_{j+2}(1)$ and the effect of an incorrect response is $u_{j+2}(0)$, where $u_{j+2}(0) = -u_{j+2}(1)$ by the constraints (7). Model (8) can be parametrized in the same way as model (2) if $u_{j+2}(1)$ is added to each parameter $u_{j+2}(x_j)$ so that the new parameter $u_{j+2}(x_j)$

+ $u_{j+2}(1)$ becomes $u_{j+2}(0) + u_{j+2}(1) = 0$ with an incorrect response and $2u_{j+2}(1)$ with a correct response. This can be done by multiplying both numerator and denonimator by

$$\exp(u_3(1)+\ldots+u_{k+2}(1)),$$

so that model (8) becomes model (2) with

$$-x_j\delta_j = u_{j+2}(x_j) + u_{j+2}(1) = x_j(2u_{j+2}(1)),$$

for all $j = 1, \ldots, k$; i.e. $\delta_j = 2u_{j+2}(0)$. This shows that the Rasch model is equivalent to the quasi-loglinear model (6).

In model (6) there is an obvious overparameterization because of the linear dependence of the item responses and the score: adding a constraint c to each of the item parameters $u_{j+2}(1)$ $(j=1,\ldots,k)$ and substracting c from $u_{j+2}(0)$ $(j=1,\ldots,k)$ to satisfy the constraints (7) is equivalent to adding
$t.c - (k-t).c = (2t-k).c$ to $u_2(t)$. This indeterminacy can be removed from model (2) by putting one linear constraint on the item parameters, e.g. by setting $u_{k+2}(x_k)$ equal to zero.

We now describe less restrictive quasi-loglinear models that can be used to detect item bias.

Quasi-Loglinear Models to Detect Item Bias.

To study item bias in a particular set of data, quasi loglinear

models may be set up that contain subgroup-dependent item parameters in addition to the parameters of the Rasch model (Rascn, 1960). The fit of these models can be compared by a likelihood ratio test with the fit of more restrictive models to test the significance of each of the subgroup-dependent item parameters. If a test yields a significant result, the item is biased. The subgroup-dependent item parameters each describe a particular type of item bias.

To detect the simplest type of bias, e.g. in item one, the model

$$(9) \qquad \ln m_{itx_1 \ldots x_k} = u + u_1(i) + u_2(t) + u_{12}(it) + \\ + u_3(x_1) + \ldots + u_{k+2}(x_k) + u_{13}(ix_1),$$

with the usual constraints (5), is compared with the loglinear Rasch model (6) to test the null nypothesis that the interaction between the subgroup and the response to item one, $u_{13}(ix_1)$ is zero. If the test is significant, it may be concluded that $u_{13}(ix_1)$ is not zero so that the difficulty of item one varies from subgroup to subgroup. The parameter $u_{13}(ix_1)$ is the change of item easyness in subgroup i and $u_3(x_1) + u_{13}(ix_1)$ is the easyness of item x_1 in subgroep i.

In model (9) a u-term is specified to test item bias for only one item. Obviously similar u-terms can be specified for two or more items if necessary. For example, comparing the loglinear Rasch model with the model:

$$(10) \qquad \ln m_{itx_1 \ldots x_k} = u + u_1(i) + u_2(t) + u_{12}(it) + u_3(x_1) + \ldots + u_{k+2}(x_k) + u_{13}(ix_1) + u_{14}(ix_2),$$

yields a simultaneous statistical test for bias in both item one and item two.

An item may be more difficult in one subgroup than another, because the item introduces some specific difficulty, e.g. reading ability, in which the members of one subgroup are generally more proficient than the members of another. If the ability to solve this difficulty varies from individual to individual within each of the subgroups and if there are two items in the test that both introduce the same difficulty we may expect these items to show an interaction that is not explained by the original latent trait.

This interaction may be investigated using the model:

$$(11) \qquad \ln m_{itx_1 \ldots x_k} = u + u_1(i) + u_2(t) + u_{12}(it) + u_3(x_1) + \ldots + u_{k+2}(x_k) + u_{13}(ix_1) + u_{14}(ix_2) + u_{34}(x_1x_2) + u_{134}(ix_1x_2)$$

which contains two u-terms, $u_{34}(x_1x_2)$ and $u_{134}(ix_1x_2)$ describing an interaction between item one and two. If $u_{134}(ix_1x_2)$ is zero but $u_{34}(x_1x_2)$ is not zero, there is a simple interaction between both items that is the same in all subgroups. If $u_{134}(ix_1x_2)$ is not zero, the interaction is different from subgroup to subgroup. This may, for example, be the case if reading ability does introduce common variance in one subgroup, but does not introduce any

variance in another subgroup, because the individuals in that subgroup are all of superior reading ability.

Comparing model (11) with the loglinear Rasch model (6) yields a test for the hypothesis that all subgroup-dependent item parameters in model (11) are simultaneously zero. If the test is significant, it may be concluded that one or more of these parameters are not zero. Comparing model (11) with model (10) yields a test for the item interaction terms alone. To test both item interaction terms $u_{34}(x_1x_2)$ and $u_{134}(ix_1x_2)$ separately, an intermediate submodel must be defined that contains $u_{34}(x_1x_1)$ but not $u_{134}(ix_1x_2)$.

---

table 2,3

Table 2 lists all relevant models (a. through e.) containing subgroup-dependent item parameters for the case of two items. Table 3 summarizes which models in Table 2 must be compared to test specific subgroup-dependent item parameters. Hypothesis 3 shows which models must be compared to test $u_{34}(x_1x_2)$ and $u_{134}(ix_1x_2)$ respectively.

Hypothesis 1-4 in Table 3 refer to what Mellenbergh (1982) has called 'uniform' item bias. It means that the bias is constant within each subgroup. With 'nonuniform' item bias (Mellenbergh, 1982) the bias of in each subgroup is dependent on the individuals ability level. Nonuniform bias may be studied witn quasi-loglinear models containing item parameters that depend both on the subgroup and the score.

Table 2 shows a series of models (f. through m.) with subgroup- and score- dependent item parameters. Since quasi-loglinear models are hierarchical, each model with a subgroup x score x item(s) interaction term must contain the corresponding subgroup x item(s) interaction term. In Table 2 all models f tnrough m contain a submodel from models a through e, which is indicated by its letter for brevity. Table 3 shows which of these models must be compared to obtain a statistical test that is sensitive to a specific type of nonuniform item bias. Note that these tests concentrate only on the nonuniformity of the bias and not on the uniform part of the bias. Therefore, if these tests are not significant, items may still be uniformly biased.

Hypothesis 5 in Table 3 concerns the simplest type of nonuniformity in item bias. If model g and f (Table 2) differ significantly, it can be concluded that the subgroup x score x item interaction $u_{123}(itx_1)$ is not zero. This nonuniformity in item bias may be expected, for example, if the difficulty of an item varies from subgroup to subgroup for low ability individuals only, which is the case if an item involves a specific skill that is not mastered by the low ability individuals of only one of the subgroups.

Hypothesis 6 (Table 3) concerns this hypothesis for two items simultaneously, whereas hypothesis 7 and 8 address the question whether item interaction is nonuniform ($u_{234}(tx_1x_2)\neq 0$) or whether subgroup differences in item interaction are nonuniform ($u_{1234}(itx_1x_2)\neq 0$). This may be called nonuniform common item bias,

where the amount of item bias that two items have in common depends on ability level. This type of item bias may occur, for example, if in only one subgroup two items introduce a common difficulty for low ability individuals but do not introduce a common difficulty for high ability subjects.

In most of the models in Table 2, the constraints are not sufficient to ensure identifiability of the model parameters. For example, the parameter $u_{23}(tx_1)$ with $t=0$ and $x_1=1$ or $t=k$ and $x_1=0$ cannot be estimated because it corresponds to structurally zero cells only. A convenient way to determine the number of estimable parameters is to determine the rank of the information matrix, which should be equal to the number of estimable parameters for a given set of data (cf. McHugh, 1956; Goodman, 1974). Baker and Nelder (1978, sec. 4.3) describe a weighted least-squares algorithm for the analysis of contingency tables, which estimates the parameters in a sequential fashion. If a parameter is linearly dependent on the preceding parameters, or if there are no observations to estimate it from, the parameter is removed from the model, thus the information matrix is of full rank.

Estimation and Testing

The kernel of the log likelihood is

$$(12) \qquad \mathcal{L} = \ln \prod_i \prod_t \prod_{x_1} \cdots \prod_{x_k} \left(m_{itx_1\ldots x_k}\right)^{f_{itx_1\ldots x_k}}$$

$$= \sum_i \sum_t \sum_{x_1} \cdots \sum_{x_k} f_{itx_1\ldots x_k} \ln m_{itx_1\ldots x_k}$$

Inserting a loglinear model for $\ln m_{itx_1\ldots x_k}$ this log likelihood yields a sum of products of model parameters (e.g. $u_3(x_1)$) with tne corresponding sufficient marginal counts (e.g. $f_{++x_1+\ldots+}$). For example, using the loglinear Rasch model (6) in (12) gives

$$(13) \qquad \mathcal{L}(\text{Rasch}) = f_{+\ldots+}u + \sum_i f_{i+\ldots+}u_1(i) + \sum_t f_{+t+\ldots+}u_2(t)$$

$$+ \sum_i \sum_t f_{it+\ldots+}u_{12}(it) + \sum_{x_1} f_{++x_1+\ldots+}u_3(x_1) \cdots$$

$$+ \sum_{x_k} f_{+\ldots+x_k} u_{k+2}(x_k).$$

where a plus sign replacing an index denotes summation over that index.

Log likelihoods of larger models (e.g. Model 9) may be obtained by adding terms (e.g. $\sum_i \sum_{x_1} f_{i+x_1+\ldots+}u_{13}(ix_1)$) to (13). If one model - say model M - is a special case of another model - say model M* - model M* may tested against model M by -2 times the natural logarithm of the likelihood ratio of both models, or equivalently, by -2 times the difference in log likelihood of both models

$$(14) \qquad G^2(M;M^*) = -2(\mathcal{L}(M)-\mathcal{L}(M^*))$$

Under the assumption of model M, G^2 is asymptotically distributed as chi-square with degrees of freedom equal to the number of estimable parameters of both models (Bishop, Fienberg & Holland, 1973, p. 525; Rao, 1965, p. 351).

An overall goodness at fit test for model M is obtained by testing it against the saturated model M* where in the expected cell counts (m) in (12) are set equal to the observed cell counts (f).

For example the Rasch model (6) is a special case of model (9). Model (9) has all parameters of the Rasch model but adds the term $u_{13}(ix_1)$. Testing model (6) against model (9) is a test for the hypothesis $u_{13}(ix_1) = 0$. If the parameter estimates of both model (6) and (9) are known, the likelihood-ratio statistic $G^2(M;M^*)$ can be calculated easily from the sufficient marginal sums corresponding to the parameters.

Maximum-likelihood estimates of the model parameters can be obtained by setting the observed marginal counts corresponding to each of the parameters equal to the corresponding expected marginal counts and solving the resulting system of equations for the parameters (Haberman, 1979, p. 448). For example, for the Rasch model the maximum-likelihood equations are

$$(15) \qquad f_{it+\ldots+} = m_{it+\ldots+} \text{ and } f_{+\ldots+x_j+\ldots+} = m_{+\ldots+x_j+\ldots+}$$

$$\text{for } i = 1, \ldots, m;\ t = 0, \ldots, k$$

$$\text{and } x_j = 0,1;\ j = 1, \ldots, k.$$

In general, for quasi-loglinear models, the maximum-likelihood equations yield no direct solution of the model parameters. The equations must be solved iteratively. Algorithms to solve the maximum-likelihood equations for quasi-loglinear models have been described by Goodman and Fay (1974: ECTA) and Baker and Nelder (1978: GLIM). Kelderman (1983) describes a generalisation to multiway tables of an algorithm by Goodman (1964, 1968) that calculates the parameters of quasi-loglinear models without setting up the entire incomplete contingency table, so that memory space required can be modest if the number of items is not small.

An Example.

Kok (1982) studied item bias in multiplication items by experimentally varying the test takers' skill in bias factors that can be expected to be operating in differently formulated test items. In this section, some of these data are reanalyzed to illustrate the use of quasi-loglinear models for the detection of item bias.

table 4

Table 4 shows the contents of six multiplication items. In item 1 through 4 the numbers are written out in Dutch and in item 5 and 6 Roman numerals are used. The subjects were 286 Dutch undergraduates of which 144 randomly selected individuals received

a short training in Roman numerals. It can be expected that the Roman items are biased.

table 5

In Table 5 for each item the values of the likelihood ratio test and the degrees of freedom are shown for both uniform (hypothesis 1, Table 3) and nonuniform bias (hypothesis 5, Table 3). From Table 5 it is seen that item 5 and item 6 are uniformly biased. There is no nonuniform bias in this set of data. Since both item 5 and 6 are written in Roman numerals, we would expect both items to be biased by a common bias factor. To test this, hypothesis 3 and 4 of Table 3 are tested. Neither showed a significant result ($G^2(c;d)=0.2$, $DF=1$; $G^2(d;e)=1.4$, $DF=1$). We can, therefore, conclude that item 5 and item 6 are uniformly biased but not that the bias factors of both items are the same.

The model with both item 5 and 6 uniformly biased (i.e. (10)) gives a good fit to the data ($G^2=106.8$, $DF=107$). The estimates for the item parameters $u_4(x_2)$ through $u_8(x_6)$ are 0.36, 0.40, -0.51, 0.05 and 0.03 respectively for $x=1$; where the first item parameter is fixed at zero. The subgroup x item response parameters $u_{17}(ix_5)$ and $u_{18}(ix_6)$ for $x=1$ and $i=1$, the group that received a training in Roman numerals, are 0.21 and .27 respectively. That is, the items are much easier for the group that received the training in Roman numerals.

Item Bias Detection when subgroups are Unknown

When subgroup membership is unobserved the subgroup variable becomes a latent variable. The models to detect item bias then become latent-class models. For example, if the latent classes are denoted by $\omega(\omega=1,\ldots,m)$, the latent class version of model (9) becomes

$$(16) \qquad \ln m_{\omega t x_1 \ldots x_k} = u+u_1(\omega)+u_2(t)+u_{12}(\omega t)+u_3(x_1) +\ldots+u_{k+2}(x_k)+u_{13}(\omega x_1)$$

$\omega = 1,\ldots,m$; $x_1=0,1;\ldots;$ $x_k=0,1$; $t=x_1+\ldots+x_k$; with the usual constraints 5.

Model (16) describes a Rasch model in each latent class ω, where the difficulty of item 1 may be different in each latent class. The parameter $u_{13}(\omega x_1)$ describes the differences in item difficulty between the latent classes. If this parameter is not zero, item 1 is biased with respect to the latent classes.

Latent-class models have been introduced by Lazarsfeld (1950; Lazarsfeld & Henri, 1968; Goodman, 1978). At first, latent-class models assumed local independence within each latent class. Goodman (1975) introduced latent-class models where the observed variables form an incomplete-contingency table assuming quasi independence within each latent class. Finally, Haberman (1979, ch. 10) formulates a latent-class model for an incomplete table where the model is not necessarily an independence model. The model can be

any identifiable loglinear model containing unobserved categorical variables. Model (16) is a special case of Haberman's general latent class model where item 1 may have a different difficulty in each of m latent classes, where the number m of latent classes is specified by the investigator. Not all latent class versions of the models to detect item bias (Table 2) make sense, since parameters involving the latent-class variable may be wholy absorbed by lower order parameters involving observed categorical variables only. These latent-class parameters are then redundant and not identifiable. This holds true for most models for nonuniform-item bias.

For example, consider the latent class version of model g Table 2:

$$(17) \qquad \ln m_{\omega t x_1 \ldots x_k} = u + u_1(\omega) + u_2(t) + u_{12}(\omega t) + u_3(x_1) + \ldots + u_{k+2}(x_k) + u_{13}(\omega x_1) + u_{23}(t x_1) + u_{123}(\omega t x_1)$$

The expected value of the observed counts $(t, x_1, \ldots, x_k)$ are then

$$m_{+t x_1 \ldots x_k} = \exp\{u + u_2(t) + u_3(x_1) + \ldots + u_{k+2}(x_k) + u_{23}(t x_1) + g_1(t x_1)\}$$

where

$$g_1(t x_1) = \ln \sum_{\omega} \exp\{u_1(\omega) + u_{12}(\omega t) + u_{13}(\omega x_1) + u_{123}(\omega t x_1)\}$$

Now $g(tx_1)$ can be completely absorbed by u, $u_2(t)$, $u_3(x_1)$ and $u_{23}(tx_1)$ to obtain new parameters using the following reparametrisation:

$$u^* = u + \bar{g}_1(++),$$

$$u_2^*(t) = u_2(t) + \bar{g}_1(t+) - \bar{g}_1(++),$$

$$u_3^*(x_1) = u_3(x_1) + \bar{g}_1(+x_1) - \bar{g}_1(++),$$

$$u_{23}^*(tx_1) = u_{23}(tx_1) + g_1(tx_1) - \bar{g}_1(t+) - \bar{g}_1(+x_1) + \bar{g}(++),$$

where the notation $\bar{g}_1(t+)$ is used to denote an average over the subscripts replaced by a plus sign. This shows that the latent-class terms in model (17) are redundant. Consequently there is no latent-class version of test 5 of table 3. A similar argument holds for test 8; the latent class term $u_{1234}(\omega tx_1x_2)$ is adsorbed by its lower order relatives involving observed variables t, x_1 and x_2.

In the latent-class models for detecting uniform bias the latent-class parameters are not adsorbed. For example latent-class version of the model used to test one-item uniform bias (16) yields the expected values of the observed counts:

$$(18a) \qquad m_{+tx_1\ldots x_k} = \exp\{u+u_2(t)+u_3(x_1)+\ldots+u_{k+2}(x_k)+g_2(tx_1)\},$$

where

(18b) $$g_2(tx_1) = \ln \sum_{\omega} \exp\{u_1(\omega)+u_{12}(\omega t)+u_{13}(\omega x_1)\},$$

so that if we set

(19a) $$u^* = u + \bar{g}_2(++)$$

$$u_2^*(t) = u_2(t) + \bar{g}_2(t+) - \bar{g}_2(++)$$

$$u_3^*(x_1) = u_3(x_1) + \bar{g}_2(+x_1) - \bar{g}_2(++)$$

$$u_{23}^*(tx_1) = g_2(tx_1) - \bar{g}_2(t+) - \bar{g}_2(+x_1) + \bar{g}(++)$$

the model

(19b) $$\ln m_{+tx_1\ldots x_k} = u^*+u_2^*(t)+u_3^*(x_1) + u_4(x_2)+\ldots+u_{k+2}(x_k)+u_{23}^*(tx_1)$$

satisfies the usual constraints (5).

In model (18) the term $g_2(tx_1)$ is not absorbed by lower order terms. The corresponding term $u_{23}^*(tx_1)$ describes a specific interaction between the test score and item 1. From (18b) it can be seen that this parameter arises both from differences in item difficulty over latent classes ($u_{13}(\omega x_1)$) as well as differences in testscore distribution in over latent classes ($u_{12}(\omega t)$) . If one of these effects are zero, the $g_2(tx_1)$ becomes constant over one index, so that from (19a) $u_{23}^*(tx_1)$ becomes zero. For example

if $u_{13}(\omega x_1)$ becomes zero $g_2(tx_1)$ does no longer depend on x_1 so that $g_2(tx_1) = g_2(tx_1')$ for all $x_1' \neq x_1$. Consequently $g_2(+x) = \overline{g}_2(++)$ and $g_2(tx_1) = \overline{g}_2(tx_1)$ for all x so that $u_{23}^*(tx_1)$ becomes zero.

If $u_{23}^*(tx_1)$ is nonzero, the item charactic curve of item one deviates from the ICC predicted by the Rasch model. This means that deviations of the ICC's of a certain item may be explained as item bias of that item with respect to unknown subgroups. Introducing latent clases may provide an alternative to introducing additional item parameters as in the two and three parameter logistic testmodel.

The latent-class versions of the remaining models for detecting uniform bias (model c-e. Table 2) also contain non-redundant latent-class terms. Writing the models for the expected values of the counts, the latent class parameters of model c-e similarly produce terms $u_{234}(tx_1x_2)$ and lower order relative terms that are not allready specified in the observed part of the model. This means that score dependent item interaction may result from differences in item difficulty or differences in item interaction between latent subgroups.

Methods for the estimation and testing of latent-class-quasi-loglinear models differ from those for ordinary quasi-loglinear models. Since latent class membership is unobserved, the frequencies $f_{\omega tx_1 \ldots x_k}$ are not known. Consequently, the maximum-likelihood equations (e.g. $f_{\omega+x_1+} = m_{\omega+x_1+}$ for parameters involving latent classes ω (e.g. $u_{13}(\omega x)$) cannot be solved because the

frequencies are unknown. Haberman (1979, ch. 10), however, gives the a rule for the derivation of maximum likelihood estimates in latent-class models from the known frequencies $f_{+tx_1 \ldots x_k}$. It says that: "The same maximum-likelihood equations apply as in the ordinary case in which all frequency counts are directly observed, except that the unobserved counts are replaced by their estimated conditional expected values given the observed marginal totals". Under some loglinear model M (e.g. Model (16)), these estimates are

$$(20) \qquad \tilde{f}_{\omega t x_1 \ldots x_k} = E_M(f_{\omega t x_1 \ldots x_k} \mid f_{+t x_1 \ldots x_k})$$

$$= (\tilde{m}_{\omega t x_1 \ldots x_k} / \tilde{m}_{+t x_1 \ldots x_k}) f_{+t x_1 \ldots x_k}$$

$$t = x_1 + \ldots + x_k$$

For model (16) the likelihood equations would then become

$$(21) \qquad \tilde{f}_{\omega t + \ldots +} = \tilde{m}_{\omega t + \ldots +} \; , \; f_{++x_1 + \ldots +} = \tilde{m}_{++x_1 + \ldots +}$$

$$f_{+ \ldots + x_k} = \tilde{m}_{+ \ldots + x_k} \quad \text{and} \quad \tilde{f}_{\omega + x_1 + \ldots +} = \tilde{m}_{\omega + x_1 + \ldots +}$$

The estimated counts $\tilde{f}$ are obtained from (20) where the $\tilde{m}$ are described by model (16). A scoring algorithm to solve these equations has been described by Haberman (1979, p. 556). An alternative way to solve these equations, is by using the E-M algorithm (Dempster Laird & Rubin, 1977) with (20) as the expecta-

tion step and (21) as the maximization step.

Discussion

In this paper an item bias detection method is proposed that uses a Rasch latent trait as an internal criterion for ability. Latent trait parameters of the model are removed from the model by conditioning on the number right score and the quasi loglinear formulation of the model is extended with parameters that describe diffrent types of item bias. The general theory of (quasi-) loglinear models is used to obtain maximum likelihood parameter estimates and likelihood ratio tests.

Using Haberman's (1979) latent class generalisation of quasi-loglinear models it is shown that even if subgroup membership is unknown it is still possible to determine whether different individuals with the same ability level have different probabilities of a correct response on a certain item.

It is also shown that nonzero item bias parameters with respect to latent classes can alternatively be modelled as parameters that describe deviations of item difficulty in different scoregroups. This means that the item characteristic curve of that item deviates from item characteristic curve predicted by the Raschmodel. Consequently, at least part of the structure in the item responses that is explained by slope parameters in the Birnbaum model may be explained as item bias. Since item bias can be interpreted as multidimensionality, item specific slope parameters may partly be

explained as multidimensionality the item response.

The models presented in this paper have two parts: one part contains parameters describing item bias, the other part contains parameters for the Rasch measurement model. It may be objected that the Rasch model is too restrictive a model for the measurement part and that a less restrictive, possibly multidimensional model, is preferred. Two remarks in favour of the Rasch model are in order here.

Firstly, as was seen before, there is a trade-off between the complexity of the item bias part of the model and the measurement part of the model. A more complex measurement model, e.g. a model with slope parameters for the item characteristic curve, may hinder the identification of certain types of item bias. Therefore if identification of item bias is the objective and nothing is known about the right (possibly multidimensional) measurement model, a simple measurement model is to be preferred. Unlike many other item bias detection methods a check of the adequacy of the item bias detection model is available because the overall fit of the model can be tested by a chi-square test.

Secondly, in general it is more desirable to construct unidimensional than multidimensional test items because the interpretation of the responses is less ambiguous. Even if a multidimensional test or item bank is needed to cover a certain content domain it is better to construct a number of homogeneous subsets of items. In that case the models presented in this paper can be applied to short subtests. Obviously, it is more probable

that short subtests fit the Rasch model than that long subtest do. For one item the Rasch model is trivially true.

Item bias detection methods using an internal ability criterion, assume that a good measure of this criterion is available, i.e. that the item used to measure this criterion fit the measurement model. If that is not the case, particularly if one or more of these items are biased themselves, the results may be erroneous. Marco (Lord, 1980, p. 228) proposed a procedure to purify a test of biased items. The total test is analyzed, items that appear to be biased are removed and the remaining items are used as an internal ability criterion to test the bias of all the testitems one by one. Although this procedure does not escape the inherent circularity of the problem it should suffice if not too many items are biased. This procedure can also be used with the test presented in this paper where in the first phase only one item-uniform bias is tested and in the second cycle the set of unbiased items is combined with pairs of possibly biased items to use the diagnostic tests presented in this paper.

Finally it should be remarked that the item bias part of the models may be more elaborate. The models in this paper contain parameters that indicate deviations due to item bias. Kok and Mellenbergh (1985) goes further and formulates models that describe the actual processes involved in the genesis of item bias more precisely. Our models may be used to give directions as to which of Kok's models may be appropriate.

Table 1

Frequency Counts and Structural Zero's in Subgroup i x Score x Item 1 x ... x Item 3 Table.

Item Response			Score t			
			0	1	2	3
x_1	x_2	x_3				
0	0	0	f_{i0000}	-	-	-
1	0	0	-	f_{i1100}	-	-
0	1	0	-	f_{i1010}	-	-
0	0	1	-	f_{i1001}	-	-
1	1	0	-	-	f_{i2110}	-
1	0	1	-	-	f_{i2101}	-
0	1	1	-	-	f_{i2011}	-
1	1	1	-	-	-	f_{i3111}

Note. Dashes denote structurally zero cells.

Table 2

Quasi-loglinear Models for Detecting Item Bias.

Models with Subgroup-Dependent Item Parameters
a. Rasch + $u_{13}(ix_1)$
b. Rasch + $u_{14}(ix_2)$
c. Rasch + $u_{13}(ix_1) + u_{14}(ix_2)$
d. Rasch + $u_{13}(ix_1) + u_{14}(ix_2) + u_{34}(x_1x_2)$
e. Rasch + $u_{13}(ix_1) + u_{14}(ix_2) + u_{34}(x_1x_2) + u_{134}(ix_1x_2)$
Models with Subgroup and Score-Dependent Item Parameters
f. (a) + $u_{23}(tx_1)$
g. (a) + $u_{23}(tx_1) + u_{123}(itx_1)$
h. (b) + $u_{24}(tx_2)$
i. (b) + $u_{24}(tx_2) + u_{124}(itx_2)$
j. (c) + $u_{23}(tx_1) + u_{24}(tx_2)$
k. (c) + $u_{23}(tx_1) + u_{24}(tx_2) + u_{123}(itx_1) + u_{124}(itx_2)$
l. (d) + $u_{23}(tx_1) + u_{24}(tx_2) + u_{123}(itx_1) + u_{124}(itx_2) + u_{234}(tx_1x_2)$
m. (e) + $u_{23}(tx_1) + u_{24}(tx_2) + u_{123}(itx_1) + u_{124}(itx_2) + u_{234}(tx_1x_2) + u_{1234}(itx_1x_2)$

Table 3

Comparison of Quasi-loglinear Models to Test u-terms for Item Bias Hypotheses.

Hypothesis	Model Forms	Comparison of Models
Uniform Bias		
1. One item uniformly biased	$u_{13}(ix_1)$	Rasch - a
2. Two items uniformly biased	$u_{13}(ix_1)$, $u_{14}(ix_2)$	Rasch - c
3. Two items with common uniform bias:	$u_{34}(x_1x_2)$	c - d
4. Two items with common uniform bias: subgroup dependent interaction	$u_{134}(ix_1x_2)$	d - e
Nonuniform Bias		
5. One item nonuniformly biased	$u_{123}(itx_1)$	f - g
6. Two items nonuniformly biased	$u_{123}(itx_1)$, $u_{123}(itx_2)$	j - k
7. Two items with common non-uniform bias	$u_{234}(tx_1x_2)$	k - l
8. Two items with common non-uniform bias: subgroup dependent interaction	$u_{1234}(itx_1x_2)$	l - m

Table 4

Multiplication Items in Dutch and Roman Numerals (from Kok 1982)

Item	Multiplication	Contents
1	7 x 1214	zeven x twaalfhonderdveertien
2	16 x 21	zestien x eenentwintig
3	16 x 14	zestien x veertien
4	6 x 4123	zes x eenenveertighonderd-drieëntwintig
5	8 x 214	VIII x CCXIV
6	5 x 1318	V x MCCCXXVIII

Table 5

Likelihood-ratio Tests for Uniform and Nonuniform Item Bias.

Item	Uniform Bias		Nonuniform Bias	
	G_2(Rasch;a)	DF	G_2(f;g)	DF
1	1.7	1	0.9	4
2	2.4	1	3.2	4
3	3.2	1	0.8	4
4	3.5	1	3.5	4
5	4.8*	1	4.0	4
6	9.9**	1	3.5	4

* $p < .05$

** $p < .005$

References

Angoff, W.H. (1982) Use of difficulty and discirmination indines for detecting item bias. In R.A. Berk (Ed.) Handbook of methods for detecting test bias. Baltimore: John Hopkins University Press.

Angoff, W.H., & Ford, S.F. (1973) Item race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-105.

Baker, F. (1981) A Criticism of Scheuneman's item bias technique. Journal of Educational Measurement, 18, 59-62.

Baker, R.J., & Nelder, J.A. (1978) The GLIM system: Generalized linear interactive modelling. Oxford: The Numerical Algorithms Group.

Berk, R.A. (1982) Handbook of methods for detecting test bias. Baltimore: The John Hopkins University Press.

Binet, A. & Simon, T. (1916) The development of Intelligence in Childeren. Baltimore: Williams & Wilkins.

Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. (1975) Discrete multivariate analysis. Cambridge, Mass.: MIT Press.

Boldt, R.F. (1983) Status of research on item content and differential performance on tests used in higher education. Research Bulletin RR-83-3, Princeton N.J.: Educational Testing Service.

Camilli, G. (1979) A critique of the chi-square method for assessing item bias. Unpublished paper, Laboratory of Educational Research, University of Colorado, Boulder.

Cardal, C. & Coffman, w.R. (1964) A method for comparing performance of different groups on the items in a test. (RM 64-61). Princeton, N.J.: Educational Testing Service.

Cleary, T.A., & Hilton, T.L. (1968) An investigation into item bias. Educational and Psychological Measurement, 8, 61-75.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM Algorithm. J.R. Statist. Soc. B., 39, 1-38.

Donlon, T.F., Hicks, M.M. & Wallmark, M.M. (1980) Sex differences in the item responses on the Graduate Record Examination. Applied Psychological Measurement, 4, 9-20.

Dorans, N.J. (1982) Technical review of item fairness studies: 1975-1979 (SR-82-90). Princeton N.J.: Educational Testing Service.

Durovic, J. (1975) Definitions of test bias: A taxonomy and an illustration of an alternative model. Unpublished doctoral dissertation, State University of New York at Albany.

Eells, K., Davis, A., Havighurst, R.J., Herrick, V.E. & Tyler, R.N. (1951) Intelligence and Cultural Differences. Chicago: University of Chicago Press.

Fienberg, S.E. (1972) The analysis of incomplete multi-way contingency tables. Biometrics, 28, 177-202. Corrig. 1972, 29, 829.

van der Flier, H., Mellenbergh, G.J., Adèr, H.J. & Wijn, M. (1984) An iterative item bias detection method. Journal of Educational Measurement, 21, 131-145.

Goodman, L.A. (1964) A short computer program for the analysis of transaction flows. Behavioral Science, 9, 176-186.

Goodman, L.A. (1968) The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing entries. Journal of the American Statistical Association, 63, 1091-1131.

Goodman, L.A. (1974) Exploratory latent structure analysis. Biometrika, 61, 215-231.

Goodman, L.A. (1975) A new model for scaling response patterns: An application of the quasi-independence concept. Journal of the American Statistical Association, 70, 755-768.

Goodman, L.A. (1978) Analyzing qualitative/categorical data: Loglinear models and latent structure analysis. London: Addison Wesley.

Goodman, L.A. & Fay, R. (1974) ECTA program, description for users. Chicago: Department of Statistics University of Chicago.

Haberman, S.J. (1979) Analysis of qualitative data: New developments, Vol. 2. New York: Academic Press.

Hoepfner, R., & Strickland, G.P. (1972) Investigating test bias. Los Angeles: Center for the Study of Evaluation, University of California.

Hunter, J.E. (1975) A critical analysis of the use of item means and item-test correlations to determine the presence of content bias in achievement test items. Paper presented at the National Institute of Education conference on Test Bias, Annapolis, MD.

Ironson, G.H. (1982) Use of chi-square and latent trait approaches for detecting item bias. In R.A. Berk Handbook of methods for detecting item bias. Baltimore: The John Hopkins University Press.

Jensen, A.R. (1973) An examination of culture bias in the Wonderlic Personnel Test. Arlington, Va.: ERIC Clearinghouse. (ERIC Document Reproduction Service ED 086 726)

Jensen, A.R. (1980) Bias in mental testing. London: Menthuen.

Kelderman, H. (1983) Generalized loglinear Rasch Models. Paper presented at the third European Meeting of the Psychometric Society, Jouy-en-Jossas, France.

Kelderman, H. (1984) Loglinear Rasch Model Tests. Psychometrika, 49, 223-245.

Kok, F.G. (1982) Het partijdige item. [The biased item.] Psychologisch laboratorium, University of Amsterdam.

Kok, F.G., & Mellenbergh, G.J. (1985) A mathematical model for item bias and a definition of bias effect size, paper presented at the Fourth Meeting of the Psychometric Society, Cambridge, Great Britain, July.

Kok, F.G., Mellenbergh, G.J. & van der Flier, H. (1985) An iterative procedure for detecting biased items. To appear.

Lazersfeld, P.F. (1950) The interpretation and computation of some latent structures. In Samuel A. Stouffer et al. (Eds.). Measurement and prediction in World War II, Vol. 4. Princeton: Princeton University Press.

Lazersfeld, P.F., & Henry, N.W. (1968). Latent structure analysis, Boston: Houghton-Miffin.

Lord, F.M. (1980) Applications of item response theory to practical testing problems. Hillsdale New Yersey: Lawrence Erlbaum.

McHugh, R.B. (1956) Efficient estimation and local identification in latent class analysis. Psychometrika, 21, 331-347.

Mellenbergh, G.J. (1982) Contingency table models for assessing item bias. Journal of Educational Statistics, 7, 105-118.

Muthen, B. & Lehman, J. (1985) Multiple group IRT modelling: Applications to item bias analysis. Journal of Educational Statistics, 10, 133-142.

Nungester, R.J. (1977) An empirical examination of three models of item bias. (Doctoral dissertation Florida State University, 1977). Dissertation Abstracts International, 38, 2726 A (University Microfilms No. 77-24, 289).

Osterlind, S.J. (1983) Test item bias. Beverly Hills: Sage.

Petersen, N.S. & Novick. M.R. (1976). An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 1976, 3-29.

Quine, M.D. & Robinson, J. (1985) Efficiencies of chi-square and likelihood ratio goodness-of-fit tests. The Annals of Statistics, 13, 727-742.

Rao, C.R. (1965) Linear statistical inference and its applications. New York: Wiley.

Rasch, G. (1960) Probabilistic models for some intelligence and attainment tests. Copenhagen: Paedagogiske Institut.

Rasch, G. (1966a) An individualistic approach to item analysis. In P.F. Lazersfeld & N.W. Henry (Eds.), Readings in Mathematical Social Science. Cambridge, Mass.: MIT Press, 89-107.

Rasch, G. (1966b) An item analysis that takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 19, 49-57.

Rudner, L.M., Getson, P.R. & Knight, D.L. (1980) Biased item detection techniques. Journal of Educational Statistics, 5, 213-233.

Scheunemann, J. (1979) A method of assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Shepard, L.A., Camilli, G., Averill, M. (1981) Comparison of procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 6, 317-377.

Shepard, L.A., Camilli, G. & Williams, D.M. (1985) Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 77-105.

Thurstone, L.L. (1925) A method of scaling psychological and educational tests. Journal of Educational Psychology, 16, 433-451.

Wright, B.D., Mead, R.J. & Draba, R. Detecting and correcting test item bias with a logistic response model (RM 22). Statistical Laboratory, Department of Education, University of Chicago.

Titles of Recent Research Reports

RR-86-1 W.J. van der Linden, The use of test scores for classification decisions with thresnold utility

RR-86-2 H. Kelderman, Item bias detection using the loglinear Rasch model: Observed and unobserved subgroups

RR-86-3 E. Boekkooi-Timminga, Simultaneous test construction by zero-one programming

RR-86-4 W.J. van der Linden, & E. Boekkooi-Timminga, A zero-one programming approach to Gulliksen's random matched subtests method

RR-86-5 E. van der Burg, J. de Leeuw, & R. Verdegaal, Homogeneity analysis with k sets of variables: An altenating least squares method witn optimal scaling features

RR-86-6 W.J. van der Linden, & T.J.H.M. Eggen, An empirical Bayes approach to item banking

RR-86-7 E. Boekkooi-Timminga, Algorithms for the construction of parallel tests by zero-one programming

RR-86-8 T.J.H.M. Eggen, & W.J. van der Linden, The use of models for paired comparisons with ties

A publication by
the Department of Education
of the University of Twente,
P.O. Box 217,
7500 AE Enschede,
The Netherlands

Department of Education